CS 441 Distributed Systems in Cloud Computing
Homework 2

Raunak Kumar Singh 675967946

Introduction

This project utilizes Apache Spark on a distributed cloud environment to parallelize computations involved in training a neural network on pre-trained Word2Vec embeddings from Assignment 1. Our goal is to predict the next word embedding in a sequence using a sliding window approach and a neural network model.

Input

The program takes as input the Word2Vec embeddings from Assignment 1. The embeddings have a dimension of 100 (E = 100). These embeddings serve as the basis for creating sliding windows to be used as input for neural network training.

Sliding Window Processing

Using Spark's distributed processing capabilities, we create sliding windows from the 30,000 vector embeddings. The sequence length for each window is set to 15 (L = 15), resulting in a tensor of dimensions (2950, 15, 100).

Neural Network Training

The output tensor from the sliding window process is then passed to a neural network for training. This computation is parallelized across the Spark cluster, allowing for efficient processing of large datasets. The resulting trained model (trained_model.zip) is saved to the root directory of the project.

Embedding Prediction

In the final step, the trained model takes a series of vector embeddings as input and predicts the next word embedding, simulating functionality of a large language model. The model's prediction outputs a vector embedding for the subsequent word in the sequence.

Baseline Submission

For baseline evaluation, we logged the following statistics in statistics.txt:

- **Training Loss and Accuracy**: Measured at each epoch to track model performance.
- **Gradient Norm**: Recorded to monitor stability and convergence of gradient-based training.
- **Learning Rate**: Adjusted dynamically across epochs.
- **Memory Usage**: Tracked to ensure efficient resource utilization.
- **Time per Epoch/Iteration**: Used to analyze computational efficiency.

- **Data Shuffling and Partitioning Statistics**: Provided by Spark’s UI (attached as part of this submission).
- **Spark-Specific Metrics**: Metrics such as job execution times and resource usage.

Training statistics

```
Sliding Windows Tensor Shape: 4777, 10, 100
Completed epoch 1, stats for it:
Gradient L2 Norm: 130.68594949176202
Epoch 1 - MSE: 0.07805169894116183, MAE: 0.2192980885173118, RMSE:
0.27937734149562277, accuracy: 0.7116347274143191
Learning Rate: 0.01
Epoch time: 445531 ms
Completed epoch 1
Completed epoch 2, stats for it:
Gradient L2 Norm: 264.6674682173986
Epoch 2 - MSE: 0.04127751126757507, MAE: 0.15985420637376319, RMSE:
0.20316867688591927, accuracy: 0.8948025790426036
Learning Rate: 0.009900990099009901
Epoch time: 540727 ms
Completed epoch 2
Completed epoch 3, stats for it:
Gradient L2 Norm: 246.82770399038228
Epoch 3 - MSE: 0.005002874384750112, MAE: 0.06266227575330775, RMSE:
0.0707310001678904, accuracy: 0.9938858769705953
Learning Rate: 0.00980392156862745
Epoch time: 425879 ms
Completed epoch 3
Completed epoch 4, stats for it:
Gradient L2 Norm: 56.697059330603885
Epoch 4 - MSE: 6.233865190737046E-4, MAE: 0.01976328910164567, RMSE:
0.02496770952798243, accuracy: 0.9984031681151525
Learning Rate: 0.009708737864077669
Epoch time: 400601 ms
Completed epoch 4
Completed epoch 5, stats for it:
```

```
Gradient L2 Norm: 73.63766829304676
Epoch 5 - MSE: 6.296992426250498E-4, MAE: 0.02152137151603535, RMSE: 0.02509380885049238, accuracy: 0.9991857919772423
Learning Rate: 0.009615384615384616
Epoch time: 433412 ms
Completed epoch 5
Total training time: 2246152 ms
```

Conclusion

This project demonstrates the effectiveness of using Spark for parallelizing neural network training on cloud infrastructure. By leveraging Word2Vec embeddings and sliding window sequences, we have successfully built and trained a language model capable of predicting word embeddings. The statistics and metrics captured during the training process provide insights into the model's performance, resource utilization, and Spark's computational efficiency.

Video Link

https://uic.zoom.us/rec/share/_4tPaRIRbNPCvwubO7HD-bpegQj0jpE8R6sYgf439THpS2US68bPNL3RlwnnVe0h.IsTMIJyo5yxueYKV?startTime=1730708820000